64-7

TECHNICAL BULLETIN 63-12

COMPARISON OF PREDICTIVE AND CONCURRENT VALIDITIES OF BASIC TEST BATTERY TEST SCORES

EDWARD F. ALF, JR.

OCTOBER 1963

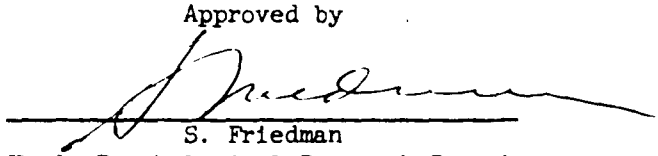
BUREAU OF NAVAL PERSONNEL

Submitted by

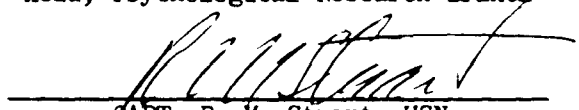
B. Rimland, Ph.D., Director, Personnel Measurement Research Department
E. E. Dudek, Ph.D., Technical Director
CDR H. B. Boaz, USN, Officer in Charge

U.S. Naval Personnel Research Activity
San Diego, California 92152

Approved by



S. Friedman
Head, Psychological Research Branch



CAPT. R. M. Stuart, USN
Director, Personnel Research Division

Bureau of Naval Personnel
Washington, D.C. 20370

BRIEF

The General Classification Test, Arithmetic Test, Mechanical Test and Clerical Test of the Basic Test Battery, which are ordinarily administered to each recruit during his fourth day of recruit training, were experimentally readministered several months later to three samples of men immediately prior to their starting training in Electrician's Mate, Hospitalman, and Interior Communications Electrician's Class "A" schools. The purpose of the experiment was to determine if the test validities would be comparable for the two administrations, since a considerable economy could be effected in evaluating new tests if the validities were found to be comparable.

Since four tests were tried at each of three schools, there were twelve comparisons between the early (predictive) and later (concurrent) validity coefficients. In each case the concurrent validity coefficient was higher, the differences ranging from .01 to .09. It was concluded that in order for a test to be considered for possible use after a concurrent tryout, it must be at least .05 to .09 more valid than the operationally given (predictive) tests. Possible explanations for these findings are discussed.

CONTENTS

TABLES

COMPARISON OF PREDICTIVE AND CONCURRENT VALIDITIES OF BASIC TEST BATTERY TEST SCORES

A. BACKGROUND AND PURPOSE

The enlisted classification tests of the Navy Basic Test Battery (BTB) are administered on the recruits' fourth day of recruit training. In evaluating an experimental test for possible inclusion in the BTB, it is not ordinarily administered at a comparable time, i.e., the recruits' fourth day in the Navy, partly because this would make it necessary to test many more recruits than would ultimately attend the schools at which validation would take place. In addition, there would be at least an eleven week delay between the time the test was taken and the time Class "A" school training was begun. Sometimes, therefore, to conserve testing time and to accelerate the validation process, experimental classification tests are administered to enlisted men at the time they enter a particular Class "A" school, and are validated against final grades obtained in that school.

Although a saving in time and money can be realized through administering experimental classification tests just prior to Class "A" school training, there is on the other hand a possibility that this practice may distort the validities obtained. It is possible, for example, that the validity of a test could be spuriously elevated if the test is administered to men just about to start Class "A" school training. The experimental test would in this case appear to be more valid than it actually was, and there would be a danger of adopting an experimental test at the expense of an operational test which was actually more valid. If, on the other hand, the validity of a test administered at Class "A" school were lower than the validity of the same test administered prior to recruit training, the opposite danger would exist. In order to make maximally effective use of validity information obtained from tests administered at Class "A" schools, it is necessary to know the extent and direction of any bias in the validities so obtained. The purpose of this study was to investigate the effects of the administration of experimental predictor tests at two different points of time on the validity of the tests.

B. PROCEDURE

The procedure followed in this study was to compare the validities of three tests in the BTB administered as a part of classification testing with the validities of the same tests administered upon entry into Class "A" school.

1. Predictor Tests

The following tests from the BTB were used as predictors:

a. The General Classification Test (GCT) is a 100-item test of verbal aptitude consisting of sentence completion and verbal analogy items (5). A single Navy Standard Score (NSS), having a mean of 50 and a standard deviation of 10, was used.

b. The Arithmetic Test (ARI) consists of two separately timed subtests (6). These are a 20-item Arithmetic Computation subtest, which provides a measure of speed and accuracy in performing elementary computations, and a 30-item Arithmetic Reasoning subtest, which provides a measure of ability to solve verbally presented quantitative problems. A total score in NSS form was used.

c. The Mechanical Test (MECH) consists of two separately timed 50-item subtests: A Mechanical Comprehension subtest, and a Tool Knowledge subtest (8). A total score in NSS form was used.

2. Predictive and Concurrent Administrations

The same forms of the above tests were administered twice to all subjects: Once during classification testing during their fourth day of training and once just prior to their entry into Class "A" school training. The regular administration during classification testing will be referred to as the "predictive" administration of the tests, and the experimental administration just prior to the beginning of Class "A" school training will be referred to as the "concurrent" administration of the tests. Similarly, "predictive validities" and "predictive means" will refer to the means and validities of the BTB tests based upon the first administration during classification testing, and "concurrent validities" and "concurrent means" will refer to the validities and means of the tests based upon their second administration, just before the beginning of Class "A" school training.

3. Subjects

The subjects for the present study comprised 267 Electrician's Mate (EM), 336 Hospitalman (HM) and 266 Interior Communications Electrician (IC) Class "A" school non-fleet trainees who entered school from September through November, 1961. All schools were located in San Diego.

4. Criterion Data

Final school grade obtained in each training program constituted the criterion. Academic drops were not included in the analysis.

C. STATISTICAL ANALYSIS

Means, standard deviations, and validities against final school grade were obtained for both administrations of the GCT, ARI and MECH for each school sample. Intercorrelations among the predictors were also obtained. Average correlations for both administrations of GCT, ARI and MECH were computed using Fisher's *r* to *z*-transformation. The significance of the difference between the predictive and concurrent validity for each test in each sample was obtained using a *t*-test for differences between correlated correlations (4, p. 148). The significance of the difference between the predictive and concurrent validity for each test, averaged over all three samples, was obtained using the *z*-test described by Winer (9, p. 44). The significance of the difference between the predictive and concurrent means and standard deviations for each test in each sample was determined using the tests for differences between correlated measures.

D. RESULTS

1. Test Validities

The validities for all tests are presented in Table 1. It can be seen that the predictive validity for each test is lower than its concurrent validity in every sample. The only two significant differences, however, are for GCT and ARI in the HM sample, both of which are significant at the .05 level ($t = 2.43$ and 2.46 respectively).

TABLE 1

Comparison of Validity Coefficients

School	N	GCT			ARI			MECH		
		Pred.	Conc.	Diff.	Pred.	Conc.	Diff.	Pred.	Conc.	Diff.
EM	267	.33	.39	.06	.19	.28	.09	.21	.27	.06
HM	336	.21	.27	.06*	.10	.17	.07*	.05	.06	.01
IC	266	.39	.40	.01	.24	.28	.04	.25	.27	.02
Mean Correlation		.30	.34	.04*	.17	.24	.07**	.16	.19	.03

Note.--

*The difference is significant at the .05 level.

**The difference is significant at the .01 level.

The mean differences across schools between the predictive and concurrent validities for GCT, ARI and MECH are also given in Table 1. These mean differences are .04, .07, and .03, respectively. The differences between the mean predictive and mean concurrent validities are significant at the .05 level for GCT ($z = 2.56$) and at the .01 level for ARI ($z = 3.09$). The mean difference for MECH was not significant ($z = 1.39$).

Examination of the validities in Table 1 reveals that the magnitude of the difference between the predictive and concurrent validity is independent of the magnitudes of the validities involved, and does not appear to be related to whether or not the tests in question were used as selectors for the schools.[1]

2. Means

The means and standard deviations for all tests are presented in Table 2. The concurrent mean for each test is higher than its predictive mean in every sample. The mean differences are all statistically significant at the .01 level of confidence.

The obtained mean differences average 3.0 points for GCT, 1.7 points for ARI and 3.1 points for MECH, with an average mean difference of 2.6 points.

3. Standard Deviations

The predictive and concurrent test standard deviations are given in Table 2. In the EM sample, the predictive standard deviations are significantly larger than the concurrent standard deviations at the .01 level for GCT, ARI and MECH. A z-test combining all three schools shows the difference between predictive and concurrent standard deviations to be significant at the .01 level for GCT ($z = 2.88$), at the .05 level for ARI ($z = 2.40$), and not significant for MECH ($z = 1.80$).

4. Predictor and Criterion Intercorrelations

The intercorrelations among the predictors and criteria for the EM, HM and IC samples are presented in Tables 3, 4 and 5, respectively in the appendix. These intercorrelations were used in computing tests of significance, and are included only for reference purposes.

---

[1] Selection requirements for HM school were a minimum total of 100 on GCT + ARI. For EM and IC schools, ARI + MECH must equal at least 105, or at least 100 and a minimum of 55 on the Electronic Technician Selection Test.

TABLE 2

Means and Standard Deviations of Test Scores

School	N	GCT Pred.	GCT Conc.	GCT Diff.	ARI Pred.	ARI Conc.	ARI Diff.	MECH Pred.	MECH Conc.	MECH Diff.
					Means					
EM	267	56.65	59.88	3.23	56.61	58.29	1.68	55.45	58.77	3.32
HM	336	56.48	59.37	2.89	53.97	55.93	1.96	48.46	51.38	2.92
IC	266	56.86	59.91	3.05	56.83	58.29	1.46	55.66	58.82	3.16
Average difference				3.0			1.7			3.1
					Standard Deviations					
EM	267	7.60	6.60	-1.00*	6.59	5.36	-1.23*	6.97	6.10	-.87*
HM	336	7.41	7.34	-.07	6.71	6.97	.26	7.53	7.61	.08
IC	266	6.76	6.59	-.17	5.60	5.37	-.23	6.09	6.06	-.03

Notes.--

All mean differences in Table 2 are significant at the .01 level.

*The difference between the designated predictive and concurrent standard deviations is significant at the .01 level.

E. DISCUSSION

In the present samples, differences between predictive and concurrent validities ranged from .01 to .09, with an average difference of about .05. Consequently, an investigator would be well advised to question the apparent superiority of an experimental test, even if it seems, on the basis of concurrent validation, that its validity is as much as .09 higher than the predictive validity of its operational counterpart.

These results are somewhat different from the results obtained by Satter and Frederiksen (7) and Frederiksen (1), who found no significant difference between the validity of a test administered during recruit classification testing and the validity of the same test administered at the end of Class "A" school training. Their study differed from the present study in a number of respects,

however, both in design and aim. The Satter and Frederiksen design used a cross-sectional rather than the longitudinal design employed in the present study (testing two groups rather than testing the same group twice). Their design raises unresolved problems in the matching of groups with regard to predictability. The interested reader will find a condensed version of their study in an article appearing in Educational and Psychological Measurement (2).

In the present study, the same tests were administered twice to the same individuals; once predictively, and once concurrently. The apparent superiority of the concurrently administered tests could be a function, in part at least, of practice effects. These practice effects should be minimal, however, inasmuch as three or more months elapsed between test administrations.

Similarly, practice effects could contribute to the obtained differences in score distributions between predictive and concurrent administrations of the tests. Despite the effect of practice upon the mean score, the data suggest that the increment in validity of the tests is not entirely attributable to practice effects. The data reveal that ARI, which shows the greatest difference in validity, is the test which shows the smallest difference in mean between the two administrations.

An alternative explanation, more consistent with the data, is that it is the time of testing which is important; the test performance of some recruits may be unduly lowered by the excitement of their first days in the Navy. Presumably the later testing took place at a time when these recruits were more relaxed and had attained a more stable adjustment to the Navy environment. In support of this explanation it is noted that, in general, the means are higher, the standard deviations are smaller, and the validities are higher for the concurrent tests than for the predictive tests. These combined factors indicate that there was a greater tendency for those scoring low on the first testing to raise their scores than for those scoring high on the first testing to raise their scores. An earlier study comparing BTB scores for third and ninth day testing of recruits attributed the higher ninth day scores to acclimatization, but validities were not available for the groups tested (3).

The present findings tend to suggest that differences between predictive and concurrent validities and means might well be a function of the type of test being used. A more thorough knowledge is needed of the degree to which different types of tests are affected by differences in testing times and conditions. If the suggestion contained in the data is confirmed that practice effect enhances validity, further research seems indicated on how best to capitalize on this finding. For example, the effects of including more practice items in a test could be evaluated, if subsequent research were to confirm that it is practice which confers an increase in validity. If, on the other hand, it is determined that "adaptation" or "acclimatization" is responsible for higher validities, consideration should

be given to the possibility of testing recruits later in training than is done at present. Further analysis of the data gathered in the present study is being undertaken to evaluate several of the above possibilities.

F. CONCLUSIONS AND RECOMMENDATIONS

Tests given immediately prior to the beginning of Class "A" school training (concurrent testing) showed a gain in validity of .01 to .09 (average = .05) over the same tests given about three months earlier (predictive testing).

While the present design does not permit differentially assessing the degree to which differences in validity may be due to practice effects or to differences in time of test administration, the findings suggest that a conservative attitude is desirable in deciding whether or not to replace an operational test with an experimental one. More specifically, experimental tests validated concurrently should show an increment in validity of at least .05 to .09 over operational tests before they are considered for adoption.

REFERENCES

1. Frederiksen, N. A further study of the validity of The Arithmetical Computation Test. Office of Scientific Research and Development. Report No. 5302. College Entrance Examination Board, Princeton, New Jersey, July, 1945.

2. Frederiksen, N., & Satter, G. A. The construction and validation of an arithmetical computation test. Educ. psychol. Measmt, 1953, 13, No. 2, pp 209-227.

3. Gordon, L. V., & Alf, E. F. Acclimatization and aptitude test performance. Educ. psychol. Measmt, 1960, 20, pp. 333-337.

4. McNemar, Q. Psychological statistics. (2nd. ed.) New York: Wiley, 1955.

5. Rimland, B. The development and standardization of Form 6 of the Navy General Classification Test. Bureau of Naval Personnel Technical Bulletin 58-4, August 1958.

6. Rimland, B. The development and standardization of Form 6 of the Navy Arithmetic Test. Bureau of Naval Personnel Technical Bulletin 58-5, August 1958.

7. Satter, G. A., & Frederiksen, N. The construction and validation of an arithmetical computation test. Office of Scientific Research and Development. Report No. 4556. College Entrance Examination Board, Princeton, New Jersey, January, 1945.

8. Swanson, L. The development and standardization of Form 6 of the Navy Mechanical Test. Bureau of Naval Personnel Technical Bulletin 58-6, August 1958.

9. Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

APPENDIX

TABLE 3

Predictor and Criterion Intercorrelations
For the EM Class "A" School Sample
(N = 267)

	1	2	3	4	5	6	7
	Concurrent			Predictive			
	GCT	ARI	MECH	GCT	ARI	MECH	CRITERION
1. GCT		.32	.24	.80	.26	.22	.39
2. ARI			-.02	.28	.66	-.03	.28
3. MECH				.26	-.02	.77	.27
4. GCT					.49	.39	.33
5. ARI						.23	.19
6. MECH							.21
7. CRITERION							

TABLE 4

Predictor and Criterion Intercorrelations
For the HM Class "A" School Sample
(N = 336)

	1	2	3	4	5	6	7
	Concurrent			Predictive			
	GCT	ARI	MECH	GCT	ARI	MECH	CRITERION
1. GCT		.51	.23	.89	.46	.22	.53
2. ARI			.17	.48	.86	.16	.50
3. MECH				.20	.16	.87	.12
4. GCT					.51	.23	.51
5. ARI						.17	.44
6. MECH							.10
7. CRITERION							

APPENDIX

TABLE 5

Predictor and Criterion Intercorrelations
For the IC Class "A" School Sample
(N = 266)

	1	2	3	4	5	6	7
	Concurrent			Predictive			
	GCT	ARI	MECH	GCT	ARI	MECH	CRITERION
1. GCT		.33	.23	.86	.25	.21	.40
2. ARI			-.02	.32	.78	-.03	.28
3. MECH				.22	-.10	.82	.27
4. GCT					.33	.22	.39
5. ARI						-.04	.24
6. MECH							.25
7. CRITERION							